International Journal of Approximate Reasoning 49 (2008) 212–233

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Bayesian learning for a class of priors with prescribed marginals

Hermann Held [a,*], Thomas Augustin [b], Elmar Kriegler [a,c]

[a] Potsdam Institute for Climate Impact Research, P.O. Box 60 12 03, D-14412 Potsdam, Germany
[b] Department of Statistics, University of Munich, Ludwigstr. 33, D-80539 Munich, Germany
[c] Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA 15213, USA

ARTICLE INFO

Article history:
Received 20 February 2007
Received in revised form 18 March 2008
Accepted 18 March 2008
Available online 6 April 2008

MSC:
62A01
62F15
62F35
68T37
86A04

Keywords:
Bayesian updating
Generalized Bayes rule
Imprecise probability
Robust Bayesian approach
Maximum likelihood update
Modeling expert opinions
Prescribed marginals
Probability of ruin
Unknown correlation structure

ABSTRACT

We discuss three learning rules for generalized Bayesian updating of an imprecise probability: (a modified version of) the generalized Bayes' rule, the maximum likelihood update rule (after Gilboa and Schmeidler) and a newly developed hybrid rule. We investigate the general methodology for a special class of multivariate probability measures with prescribed marginals but arbitrary correlation structure. Both the choice and analysis of this class are motivated by expert interviews that we conducted with modelers in the field of climatic change.

We argue that both updating rules from the literature have strong limitations, the generalized Bayes' rule is too conservative, i.e., too inclusive, while the maximum likelihood update rule being too exclusive, adding spurious information. As a powerful extension we introduce a new rule for Bayesian updating of an imprecise measure: a "weighted likelihood update method," which bases Bayesian updating on the whole set of priors but weights the influence of its members. We study the different rules in the case of bivariate Gaussian priors. Our investigation shows that the new rule combines certain attractive features of the generalized Bayes' rule and the maximum likelihood update rule. In this article we aim at highlighting the sequence of not yet fully resolved statistical issues a practitioner on complex mechanistic models would face when updating imprecise prior knowledge.

1. Introduction and background

Imprecise probability [12,15,45,47,49,50] provides a means to express ambiguity (non-stochastic uncertainty) in an appropriate way and is equally a powerful methodology for generalizing the classical calculus of probability to handle the multi-dimensional nature of uncertainty [30, p. 1], which in particular arises when expert opinions have to be modeled.

The most common – closely related – mathematical tools are non-additive set functions and sets of classical probabilities. We rely strictly on the latter throughout this paper.

* Corresponding author.
E-mail addresses: held@pik-potsdam.de, hermann.held@pik-potsdam.de (H. Held), thomas@stat.uni-muenchen.de (T. Augustin), elmar@cmu.edu (E. Kriegler).

0888-613X/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2008.03.018

1.1. Updating rules

Since expert opinions are typically to be understood as a sort of generalized subjective prior probability, it is a natural aim to generalize the classical Bayesian approach of learning. The latter utilizes Bayes' theorem to update prior knowledge, formulated by prior probabilities over parameters, in the light of data to obtain posterior probabilities on the parameters, which then completely describe the state of knowledge after having seen the data.

One way to extend this learning scheme to imprecise probabilities represented by sets of prior probabilities suggests itself: the element-by-element approach, where each precise prior is updated to the corresponding posterior, and then the whole set of posterior probabilities is used for the subsequent analysis. This way to proceed seems to be self-evident for most robust Bayesians (e.g., [38]). Moreover it can be corroborated by strong internal consistency arguments: Walley [45] neatly justifies his closely related *generalized Bayes rule* (abbreviated as GBR-W hereafter) by deriving it from general coherence axioms within his framework. Consequently GBR-W has become one of the central tools in the theory of imprecise probability.

However, one has to be aware that frequentist and decision theoretic optimality properties inherent in the classical Bayesian approach do not extend to the imprecise probability setting, in particular decision functions derived from the GBR-W are not risk-minimizing in data-based decision problems [1].

The main reservation towards the practical use of GBR-W, however, comes from the applied point of view: The range of the resulting posterior probabilities may turn out to be quite large, and sometimes even too large to allow for any statement that is non-trivial from the practical point of view. This is a very subtle issue, since a wide posterior range is not always undesirable: [45, p. 5] explicitly names the possibility to recognize prior-data conflict in wide posterior intervals as one of the main motivations for imprecise probabilities, and [40] give even arguments in favor of the possibility of the phenomenon of dilation, where imprecision increases irrespective of the data. On the other hand, for instance in climate modeling, almost vacuous posterior intervals are often counter-intuitive, and are then perceived as simple mathematical artefacts resulting from "extreme priors" that are contained in the prior class for the sake of simplicity of mathematical treatment, but without any meaning from the substantive point of view. In those situations it is highly desirable to exclude such "meaningless priors" from the analysis. A first step towards this aim is to add some qualitative restrictions to the original class, filtering out priors of extreme shape that usually do not fit to any expert's judgment. There is a long list of possible restrictions available in the context of robust Bayesian analysis (see, e.g., the models applied in [38]); we will later on explicitly apply a gradient filter, which was also used in [46].

Another, much more fundamental, way of avoiding near vacuous posterior conclusions is a so-to-say dynamic revision of the underlying set of probabilities in the sense that the observed data are used for weighting the prior opinions. In practical terms this can be interpreted as the fact that after having seen the data some prior expert opinions appear to be less reliable than others; from the theoretical standpoint this means nothing less than the search for alternative learning rules. Walley himself sees plenty of room for discussion,[1] but only few alternative rules have been studied so far. These rare exceptions include [41], refining priors in the light of the data, [10,11], developing an updating procedure driven by an information measure, Gilboa's and Schmeidler's [26] *maximum likelihood updating*, which restricts consideration to those priors which will have foreseen the observations with highest predictive prior probability (see also [48, Appendix] for an argument in that direction), and a quantile-based selection of priors [27].

We will depart from the maximum likelihood updating method that is closely related to Dempster's rule of conditioning [16], and so, in this vein, any other rule of conditioning could be used (see [52] for different types of conditioning in Dempster–Shafer theory, and also [19,51]) in principle, as long as one still relies on the implicit Bayesian dogma that updating is best done by conditioning.[2] Moreover, note that there is also a close relationship to more traditional Bayesian statistics. In terms of [4, p. 99] every such prior is a *type II maximum likelihood prior*, and so an imprecise posterior derived from the maximum likelihood update method can be interpreted as the envelope of all the posteriors arising from type II maximum likelihood priors.[3]

In investigating the maximum likelihood update rule, we gain the impression that it is excessively prone to adding spurious information in an overoptimistic way. As a compromise we suggest a new hybrid rule, called *weighted likelihood update rule*, that weights the original elements of the prior in a data-dependent way, linearly: We divide the set of priors into different strata according to their predictive power related to the data, and define the posterior interval as the weighted sum of the posterior intervals corresponding to those strata.[4]

[1] Walley [45, p. 335] notes that although "the earlier theory of coherence suggests that [...the GBR-W] is a reasonable updating strategy, [...] there is nothing in the theory that requires You to follow either" the construction of conditional previsions through the GBR-W or the adoption of the resulting conditional prevision as the updated prevision. He also stresses (c.f. [45, p. 334]) that "there is a role for other updating strategies, not because the updated beliefs constructed through the GBR-W are unjustified, but because they are often indeterminate" and gives a list of twelve items (cf. [45, Chapter 6.11.1]) summarizing "[...] the reasons for which the GBR-W may fail to be applicable or useful as an updating strategy".

[2] [2, Section 1] stress that, at least in the predictive setting considered there, conditioning and updating have to be well distinguished.

[3] This point of view could be very helpful for practical computations, since there are techniques based on the EM-algorithm available for calculating type II maximum likelihood priors, e.g., [4, p. 100].

[4] In the light of the concept of type II maximum likelihood an alternative approach to refine the prior should again be briefly mentioned here: [14] consider conditional Γ-minimaxity actions among the set of all priors "close" to the type II maximum likelihood prior.

1.2. Probabilities with prescribed marginals

Here we investigate three different ways of Bayesian updating (pointwise updating of the prior class, closely related to GBR-W, maximum likelihood and weighted likelihood updating) for a particular imprecise probability model. The model is described by a class of multivariate probabilities with prescribed marginals. It shall represent epistemic uncertainty on a multi-dimensional parameter space where an expert is able to specify any of the marginals in terms of probability measures, but refuses to deliver any further information, in particular not on the correlation structure among the parameters. A couple of authors have investigated such classes [6,17,22,27,32–34,42]. Some of the properties of those classes are also resembled by the multivariate possibility measure [18] introduced in [29]. For that possibility measure, instead of an explicit correlation structure, [29] prescribe spherical symmetry in a rather heuristic manner.

However this article goes one step further and compares posterior properties of the class with prescribed marginals for the competing updating rules generalizing Bayesian inference. This distinguishes it from earlier analyses, e.g., [32], who investigate posterior properties derived from the generalized Bayes rule only.

Our interest in the class of prior probabilities with prescribed marginals arose when we observed that climate model developers or users frequently claimed a lot of knowledge on individual model parameters related to specific physical processes, but felt much less able to give any prior knowledge on the way the parameters must interact in order to obtain a reasonable model climate state. The situation is similar for other models used in the climatic change assessment. We based our impression on more objective grounds by distributing a questionnaire to half a dozen model users (see Section 3.1). In this paper, we would like to analyze the implications of the apparent difference in availability of "knowledge on marginals" versus "knowledge on correlations" and choose a precise probability measure for the marginals and complete ignorance about the correlation structure.

It is our impression that an investigation of this type is desperately needed as in the climate modeling community – as well as in many other research communities – the issue of prior knowledge on parameter correlations is one of the most neglected issues, typically being represented in a naive way by uncorrelated measures. The situation is different as far as univariate uncertainties about single parameters are concerned. There are examples in the literature that include something close to robust Bayesian analysis in a rudimentary manner (in [21] two sorts of priors are investigated) or even explicit treatment in terms of probability ratio models [43] or probability boxes [31].

However, the frequently made silent assumption that an expert uninformed about correlations is best represented by uncorrelated parameters, seems to us to mimic the "objective Bayesians'" assumption that the situation of complete ignorance on a single parameter is best represented by a non-informative prior. We follow the quite contrary assumption of Walley [45, Chapter 5.5] that there is no such thing as non-informative priors and that situations of ignorance must be captured by imprecise models.[5] Therefore we proceed in setting up an imprecise model for correlations.

While a class of priors with prescribed marginals seems to be what follows directly from the interviews with experts, [34, Section 3.3] calls the imprecise posterior (after updating with GBR-W) of such classes "useless" (i.e., almost non-informative). In [34, Section 3.3] a contamination class is suggested instead that is a superposition of a precise prior and the class of vacuous correlation structure we would like to investigate here. So the posterior is made more informative by returning to an "almost precise" prior. On the contrary we find it more convincing to exclude the most bizarre elements of the original class (Section 2.3) and to search for alternative updating rules (Section 2.1).

It is apparent that the prior correlation structure will have a strong influence on the result of generalized Bayesian updating, in particular in high dimensions. For example, for a non-informative likelihood and identical Gaussian marginals, the standard deviation of the posterior will scale with $\sim \sqrt{n}$ for the uncorrelated case, but with $\sim n$ for the perfectly correlated case (n denoting the number of parameters). This article highlights the interplay of such effects with the mechanisms of inclusion and exclusion as manifest in the different updating rules, for the analytically most transparent cases. Finally, one could ask what the effects of imperfections in the elicited prior were, and add another layer of uncertainty in the sense of robust Bayesian analysis [4]. However, such analysis is clearly beyond the scope of this article. Here we would like to study the consequences of the main imprecision in the prior as elicited under the three updating rules.

In more detail, the present article is organized as follows: Section 2 is devoted to the general development and discussion of learning rules. We start with two prominent generalizations of Bayes' formula for imprecise priors, then introduce a new generalization and also make some suggestions on how to deal with overly inclusive prior classes. We then investigate the learning rules in more detail in a special situation. Section 3 describes the underlying prior model, including our expert elicitation motivating it. The results are presented in Section 4. The newly introduced weighted likelihood updating rule appears as more convincing than the overly exclusive maximum likelihood update rule proposed by Gilboa and Schmeidler, while at the same time it generates much more conclusive statements than can be inferred from the very conservative GBR, a modified version of GBR-W. Finally, in Section 5 we summarize our results and outline needs for future work.

[5] A bit more flexibility is achieved by specifying copulas [35], to which, however, similar counter-arguments also apply, since the dependence structure is still specified more precisely than can be derived honestly.

2. Generalizations of Bayes' formula

In this section we start with the basic form of three updating rules and then discuss several ways to reduce overly inclusive prior sets. All the rules generalize traditional Bayesian learning (updating) on a possibly vector-valued parameter x from data y, given a single prior probability measure P, by Bayes' rule:

$$P_{\text{post},y}(x) = \frac{P(x)L_y(x)}{\int \mathrm{d}x' P(x')L_y(x')}. \tag{1}$$

In this formula we require that x assumes values of a continuous random variable (i.e., has a density with respect to the Lebesgue measure) and $0 < \int \mathrm{d}x' P(x') \times L_y(x') < \infty$, as we will assume silently throughout the rest of the paper. $L_y(x) \equiv P(y|x)$ denotes the likelihood function for the uncertain (multivariate) parameter x. Note further that here and later in the paper we use "P" synonymously for the probability measure as well as for the accompanying density when applied to absolutely continuous probability distributions on $\mathbb{R}^n$.

2.1. Updating rules

However, since in our situation we have a class of priors rather than a single prior, Eq. (1) cannot directly be applied in the traditional way, and we have to specify appropriate generalizations. In the following we outline types of generalizations of Bayes' formula we will employ when updating a class of priors $\mathscr{P}$ rather than a single prior P. We briefly summarize two update rules discussed in the literature, the *generalized Bayes rule (GBR)* (within this article a straightforwardly modified version of the generalized Bayes rule GBR-W for convex sets of probabilities [45, Chapter 8.4.8] is used) and the *maximum likelihood update rule* (after [26]). As a third option we will introduce the *weighted likelihood update rule* that includes elements of both of the former rules.

The three updating rules work as follows:

2.1.1. GBR

(1) In view of evidence y, any element of the class of priors $\mathscr{P}$ is updated according to (standard) Bayes' rule (1).
(2) The probability of interest (e.g., of crossing a certain threshold, in our application the *probability of ruin*, Eq. (7) below) is extracted from all of the posteriors.
(3) The inf- and sup-operations are applied to find the lower and upper probability of interest. Thereby an interval, for instance, of probabilities of ruin is created.

As discussed in the Introduction, this rule is standard procedure in robust Bayesian inference (e.g., [5], (3) making up the so called "posterior imprecision" [34, Section 3.1]) and in the behavioral approach to imprecise probability initiated by Walley [45], who also justifies this way of updating by deriving it from a list of coherence axioms. In the following we will use the term "GBR" not only for updating of $\mathscr{P}$ but also as a module in other learning rules where the above three steps are applied to subsets of $\mathscr{P}$.

2.1.2. Maximum likelihood update rule [26]

(1) The subset of $\mathscr{P}$ is determined that contains all those priors that optimize the predictive prior probability (density) for the measurement y, i.e.:

$$\mathscr{P}_{\text{ML},y} := \operatorname{argmax}\left\{ P \in \mathscr{P} \,\middle|\, \int \mathrm{d}x' P(x')L_y(x') \right\}.$$

(2) GBR is applied to this subset.

The method completely disregards expert opinions that have not foreseen the measurement y with maximum probability. Gilboa and Schmeidler [26] showed that the maximum likelihood update method coincides with a generalized version of Dempster's rule of conditioning [16] (to arbitrary coherent lower probabilities) if the lower probability on the joint space $X \times Y$ of possible parameters and observations is 2-monotone. Although it is not unrealistic to discount expert opinions[6] that are at odds with observations, we find it unconvincing to dismiss *completely* opinions just because they have not foreseen the measurement with maximum probability. In particular, our discomfort refers to the exclusion of those opinions that have missed the maximum by just an infinitesimal amount.

[6] One criticism of this method points to the fact that the evidence is used twice – first for selection of priors, then for GBR. In this context the connection to type II maximum likelihood methods mentioned in the Introduction is helpful. It relates the procedure to empirical Bayes methods (e.g., [4, Section 4.5]), where such a double use of data has been justified and is widely accepted. Further discussions of this point will follow in footnote 9 and in the concluding Section.

2.1.3. The weighted likelihood update rule

For that reason, we introduce an extension of the maximum likelihood update method, the *weighted likelihood update rule*. On the one hand, we require that any of the prior opinions is considered in the updating process, a property shared with GBR, and on the other hand, as in the maximum likelihood update rule, that the predictive power of each element of $\mathscr{P}$ should play a role. Our new method involves the following steps:

(1) We decompose $\mathscr{P}$ in terms of level sets with respect to prior probability of y (Eqs. (3)–(5) below).
(2) We apply GBR to each level set.
(3) We average over GBR results while we linearly weight with respect to the prior predictive probability density of y.[7]

For a technical definition of our new method let for the moment $\mathscr{P}$ be of finite cardinality I, i.e., $\mathscr{P} = \{P_1, \ldots, P_I\}$. Let $W_y : \mathscr{P} \to \mathbb{R}_0^+$ denote the probability of y for any prior P, i.e., for given y

$$\forall_{P\in\mathscr{P}} \quad W_y(P) := \int \mathrm{d}x\, P(x)L_y(x) = \int \mathrm{d}x\, P(x)P(y \mid x), \tag{2}$$

assuming that for all $P \in \mathscr{P}$ the relation $\int \mathrm{d}x\, P(x)L(x) < \infty$ holds.

Let $\{w_{1,y}, \ldots, w_{J,y}\} := W(\mathscr{P}), J \leqslant I$ be the set of weights generated from $\mathscr{P}$, i.e., the prior probabilities of the evidence y, and $P^*_{\mathrm{post},y}(P)$ the posterior probability of the quantity of interest, given the prior $P \in \mathscr{P}$. Based on

$$\mathscr{P}_{j,y} := \{P \in \mathscr{P} \mid W_y(P) = w_{j,y}\}, \quad j = 1, \ldots, J, \tag{3}$$

we define

$$\underline{P}^*_{\mathrm{post.wm},y} := \frac{\sum_{j=1}^J w_{j,y} \cdot \inf\limits_{P\in\mathscr{P}_{j,y}} (P^*_{\mathrm{post},y}(P))}{\sum_{j=1}^J w_{j,y}}, \tag{4}$$

$$\overline{P}^*_{\mathrm{post.wm},y} := \frac{\sum_{j=1}^J w_{j,y} \cdot \sup\limits_{P\in\mathscr{P}_{j,y}} (P^*_{\mathrm{post},y}(P))}{\sum_{j=1}^J w_{j,y}}, \tag{5}$$

assuming $\sum_{j=1}^J w_{j,y} > 0$. This new method would reveal results identical to those obtained from the maximum likelihood update if $w_{1,y} \neq 0, w_{2,y}, \ldots, w_{J,y} = 0$. Furthermore, if W_y is injective, the weighted likelihood update rule (WLU) reduces to precise Bayesian updating with j as a hyperparameter and equal prior weight assigned to all elements of $\mathscr{P}$. The latter has been introduced in [39] and analyzed in [3] within the context of combining the posteriors from different experts.[8]

In Section 4.5.2 we extend WLU to a $\mathscr{P}$ of infinite cardinality. If $\mathscr{P}$ is integrable with respect to W and in the special case W_y is injective (which is not the case for the example in Section 4.5.2), WLU delivers the measure in the hyperparameter for precise Bayesian updating, in assuming W as uniformly distributed.[9]

Being lower and upper envelopes of set of precise probabilities, the lower and upper posterior probabilities derived from the maximum likelihood updating rule and the weighted maximum likelihood update rule are still separately coherent in the sense of [45, Def. 6.2.2], but they fail to satisfy the stronger condition of being coherent to the unconditonal prior probabilities in the sense of [45, Section 6.3], since they do not obey the generalized Bayes rule, which can be derived from that stronger type of coherence (see [45, p. 296, 6.3.5., Property (10) and Section 6.4]).

Finally we would like to stress that our WLU is introduced on the purely heuristic grounds outlined above. A deeper theoretical underpinning is beyond the scope of this article, however would be highly desirable. We will find that for our somewhat paradigmatic imprecise prior, WLU delivers much more intuitive results than ML, and is much more informative than GBR.

We will refer to these three generalizations of Bayes' formula also as "learning rules" in the following.

2.2. Interpretation of overly inclusive prior classes

We would like to address a further conceptual difficulty that occurs when dealing with classes of priors $\mathscr{P}_0$ assigned by some easy-to-handle mathematical model. For instance our (stylized) class of priors with prescribed marginals (see (8)) may be too inclusive for a particular application, i.e., it may contain priors that do not correspond to any reasonable expert opinion, and so $\mathscr{P}_0 \supsetneq \mathscr{P}$, where $\mathscr{P}$ describes the correct set of priors. The opinions in $\mathscr{P}_0 \setminus \mathscr{P}$ may drastically distort the inferred upper (lower) probabilities. There are two ways how to deal with such a "contaminated class" $\mathscr{P}_0$:

[7] Alternatively one could think of non-linear forms of weighting or simply applying GBR to the class of priors that are characterized by a certain minimum of weight. However, such generalizations shall be investigated elsewhere.

[8] The above weighting rule represents an "improper scoring rule" [9], p. 137, as it poses an incentive to competing experts to produce over-confident statements [3], [9], p.137. However here we do not have to deal with the "properness" of scoring as our imprecise prior is supposed to be a model for a single, self-consistent expert who is supposed to be properly elicited.

[9] This demonstrates that WLU does not involve "double counting" (see also footnote 6 on that issue) of data in the standard sense, where data may be used twice within sequential Bayesian updating. Here data are used twice, but in "orthogonal" ways. The issue of "using the evidence twice" is of concern for Gilboa's and Schmeidler's rule as well.

(1) If the main sources of such distortion are known, one simply adds a *filter* that eliminates unrealistic priors. In Section 2.3 we will suggest a catalogue of such additional filters that should be observed in standard applications. For example, we will argue that only those priors should be considered further that come with a density whose gradient does not transgress a certain norm.
(2) If the wrong elements of the prior class cannot be filtered out, one has to be aware that the three rules discussed show quite different behavior. For GBR the following convenient relationship holds, which is a direct consequence of the definition of GBR:
When using GBR, the upper (lower) probability derived from the overly inclusive class of priors $\mathscr{P}_0$ serves as an upper (lower) boundary of the upper (lower) probability derived from the correct class $\mathscr{P}$.
This property conveniently implies that – in the case of the GBR – we are always on the safe side (i.e., we do not add spurious information) when we include also those priors we are not sure about yet.
On the contrary, for the weighted as well as for maximum likelihood update method, the probability interval derived from an overly inclusive class $\mathscr{P}_0$ must not be interpreted as a outer boundary of the correct probability interval resulting from $\mathscr{P}$. Therefore the weighted or maximum likelihood update method can be used only *after* we have finally decided for any prior whether it should enter the class or not.
To illustrate this point, imagine an assembly of "experts" $E_1,\ldots,E_i,\ldots,E_M$, most of whom may in fact be charlatans, who for any $i \in \{1,\ldots,M\}$, assign 100% chance to lottery result y_i and zero to any y_j with $j \neq i$, then accidentally the corresponding opinion P_i would be highlighted if y_i was measured. So a lot of trust would be given to a potential charlatan's opinion P_i although it was just chosen for trivial reasons and not because it was characterized by higher a priori competence.

Therefore, filtering out false priors would have two effects: GBR would become more informative and we would avoid distortions in using the weighted or maximum likelihood update rule.

Note that in the present article, we purposefully suppress another layer of complication: in applications, it will generically prove very difficult to completely "purify" the set of priors from "contaminations" as mentioned above and – vice versa – to make sure that none of the "correct" precise measures was included in the set of priors. Along the lines of robust statistics one would therefore analyze the "robustness" of the (in our case imprecise) posterior along minuscule changes of the (in our case "purified" imprecise) prior. In this article we would like to highlight the effects of changing the update rules as distinct from those stemming from varying the prior. Future work will have to show how posterior robustness against minuscule changes of the imprecise prior should be operationalized and whether there is a ranking among the three update rules introduced above with respect to that kind of robustness.

2.3. Further constraints on the class of priors

Since we refocus on modeling the prior knowledge of one single expert, we ask what such an expert generically would be able to hold an informed opinion on, in order to narrow down the class to reasonable priors:

(1) The density of the priors should be unimodal [34, Section 3.2.4] for an overview of references.
(2) We assume that the typical 1D (i.e., marginal) resolution over which an expert can hold an informed opinion, reads $\mathrm{d}x_1$ (here "1" for "1D") if the typical dimension of the problem is Δx. This implies that an expert can distinguish $N_1 \approx \Delta x/\mathrm{d}x_1$ items. Our requirement is equivalent to Walley's "bounded derivative model" [46] and shall be called *gradient filter* in the following.
(3) This prescription needs to be generalized to a n-dimensional parameter space:
   (a) A possible generalization that would lead to a particularly large prior class is obtained by allowing for a resolution in terms of cubes of length $\mathrm{d}x_1$, i.e., $N_n \approx N_1^n$.
   (b) The other extreme may require that $N_n \approx N_1$.We can connect both extreme cases by
   $$N_n := N_1^{\beta n+(1-\beta)}, \tag{6}$$
   hence we construct the linear hull of the exponents of both cases, $\beta \in [0,1]$. Such a connection may turn out to be necessary as both extreme cases display dissatisfying features:
   (a) Let $\beta = 0 \Rightarrow N_n = N_1$. As $N_n = \Delta x^n/\mathrm{d}x_n^n$ (with $\mathrm{d}x_n$ denoting the length of the edge of the n-dimensional cube), we observe: $\lim_{n\to\infty}\mathrm{d}x_n = \lim_{n\to\infty}\Delta x/N_n^{1/n} = \Delta x$. This demonstrates that the expert may not have much knowledge left on the nD parameter space, measured in terms of 1D information $\mathrm{d}x_n$.
   (b) On the other hand, $\beta = 1 \Rightarrow N_n = N_1^n$ would require a prior competence of the expert, exponentially growing with dimension, that seems unrealistic as well.Both phenomena are rooted in the "curse of dimension". Hence, there is urgent need for an expert elicitation that is designed to obtain a meaningful intermediate value for β. For the time being we derive the consequences of various values for β. Once β has been decided on, the current (third) prescription on prior distributions requires that the modulus of the distribution's gradient to be smaller than $(1/\mathrm{d}x_n^n)/\mathrm{d}x = \Delta x^{-(n+1)} \cdot N_1^{\beta n+1+(1-\beta)/n}$.

2.4. Specification of the output quantity of interest

In order to keep the discussion as transparent as possible, in the following we will focus on a single ouput quantity of interest whose posterior probability shall be derived. For that we choose the *probability of ruin*

$$P^*_{\text{post},y} = \int_{x_1^*}^{\infty} \mathrm{d}x_1 \int_{-\infty}^{\infty} \mathrm{d}x_2 P_{\text{post},y}(x_1, x_2), \tag{7}$$

i.e., the probability that x_1 (e.g., climate sensitivity) is larger than a certain threshold value x_1^*.

3. Priors with given marginals

In this Section, we will specify our prior model used in the investigation later on. The model considers given marginals, but leaves the correlation structure completely open. This model has also attracted some attention outside the area of Bayesian updating (cf. [6,17,22,32,33,42]), the resulting lower and upper joint probabilities are then typically called Fréchet bounds in this setting [23].

3.1. A questionnaire on the structure of prior knowledge

3.1.1. Design of the questionnaire and choice of experts

We have claimed above that climate change modelers typically not only know more about individual parameters than about their correlations but are often even completely unable to elicitate correlation. In order to underpin this claim we developed a questionnaire on the structure of prior knowledge about model parameters. We considered users of the following models:

- The climate model of intermediate complexity CLIMBER-2 (corresponding to a system of more than 1000 ordinary differential equations) [24,37].
- The complex ocean model MOM-3 [36].
- The dynamic vegetation model LPJ [13].
- The model of endogenous economic growth MIND [20].

We asked them – among other items – whether for a given uncertain model parameter b, the expert would be willing to give probabilistic information in terms of a density function. *Any* of experts that participated in the survey would do so. Then we checked for the betting behavior on quantiles of b. We found certain discrepancies with the density function specified before that may suggest the use of imprecise measures on b. However, these aspects are not discussed here and will be published elsewhere, together with the questionnaire. In the context of this article, we would like to focus on the central question concerning the choice of our model (8):

> *How do you judge the quality of your subjective knowledge on b compared to the quality of your knowledge on correlations of b with other unknown parameters?*

3.1.2. Answers collected in the expert elicitation

Typical expert responses from the survey are quoted below:

- "For some of the parameters, I know about the sign of correlations, however my knowledge is less precise than that on individual parameters."
- "Knowledge on b is higher than knowledge on correlations with other parameters (on some specific parameters, it might be different)."
- "The parameter knowledge is relatively good but the knowledge on correlation with other parameters in some cases is only an idea."
- "I have not considered the possibility of correlations."
- "Never thought about that point."
- "It would be impossible to specify anything on correlations."
- "Absolutely no comment on correlations."

We would like to stress again that those statements were made by experts who at the same time were willing to specify prior knowledge in probabilistic terms on individual parameters! Our rather informal kind of elicitation, to a certain extent, establishes much less precise knowledge on correlations than on marginals. In case the actual numbers for specific models were to be elicited in the future, including rudimentary knowledge on correlations, we shall observe existing methodologies on multivariate elicitation [25, Section 2.3]. However, [25] stress the difficulty to elicit multivariate rather than univariate entities, an instance adding further imprecision to the representation of correlations.

3.2. Consequences for our choice of an imprecise model

Therefore we feel the urgent need to consider the somewhat extreme case of imprecise prior knowledge with prescribed marginals (i.e., knowledge on individual parameters) and fully unconstrained correlation structure. This class – already discussed in different settings, in particular, in [6,17,22,32,33,42] – is defined for the case of two marginals P_1, P_2 by

$$\mathscr{P} := \left\{ P \mid \forall_{x_1} \int \mathrm{d}x_2 P(x_1, x_2) = P_1(x_1), \forall_{x_2} \int \mathrm{d}x_1 P(x_1, x_2) = P_2(x_2) \right\}, \tag{8}$$

where P_1, P_2 are specified by the expert. Hereby for simplicity within this article we assume again continuous random variables.[10] Note that any element P of this class is by construction normalized to 1.

There are some analytical and procedural results on classes of priors with prescribed precise marginals before [22,23,35,42] and after updating [32,34]. Here we abstain from using those results since they do not apply to non-convex classes as considered here. The updating procedure of it, in particular according to our new WLU, requires some extra treatment which is easiest performed with an explicit parameterization of our prior.

4. Updating our imprecise prior

4.1. Specifying the marginals and the likelihood function

In this Section, we apply all three generalized learning rules to the class of priors constrained by Gaussian marginals. (As any non-degenerate marginal specified by an expert can be mapped onto a Gaussian through a suitable coordinate transformation, no loss of generality is introduced in this instance.) For comparison, we also present the result of a conventional Bayesian analysis where a precise Gaussian prior is derived from the marginals by assuming independence between the parameters and updated using Bayes' rule. We then repeat the analysis for strict subsets of the original class of priors. These subsets consist of those priors which have passed two versions of the gradient filter introduced in Section 2.3. Such prior classes seem to model an expert's knowledge more realistically. Finally, we summarize our results in considering the example of how the stylized insurance company may base its contracting decisions upon it.

We specify the marginal $P_1 \sim N(\mu, \sigma^2)$ as a Gaussian with mean μ and variance[11] σ^2, and take $\mu = 1/2$, $\sigma = 1/4$, $P_2 \equiv P_1$, hence we select marginals that contain ±2 standard deviations in $[0,1]$.

The most general class of priors with prescribed marginals has been defined in Eq. (8) in Section 3.2. However, for this article we sacrifice generality for an analytically elegant and transparent implementation of the otherwise intricate and potentially only numerically accessible unimodality filter. We consider the simplest non-trivial choice of the class of priors, and require that any prior should be a bivariate Gaussian. That choice ensures that any prior is unimodal.[12] Later on we will also require bounds on the gradients, thereby avoiding degenerate, essentially lower-than two-dimensional Gaussians (see Fig. 1, left and right graphs). Before that, however, we would like to study Bayesian updating on the class of bivariate Gaussians with unrestricted gradients.

For simplicity we assume further that the likelihood function L is Gaussian as well and x_1,x_2 determine its mean through a linear transformation $(x_1, x_2) \to \kappa x_1 + x_2, \kappa \in \mathbb{R}$ under the "observation" y. So we have

$$y = \kappa x_1 + x_2 + \eta, \quad \eta \sim N(0, \sigma_\eta^2) \tag{9}$$

and

$$L_y(x_1, x_2) \equiv P(y \mid x_1, x_2) := N(\kappa x_1 + x_2, \sigma_\eta^2)(y). \tag{10}$$

This likelihood could be interpreted as a climate model that from the property x_1 would predict κx_1 (with well-known κ) which in turn could in principle be compared to the noisy observation y. However modelers may know that the model could have a systematic bias x_2, hampering direct comparison with y. Such statements have actually been made on the model CLIMBER-2 (highlighted in the expert elicitation) with {prior variance$(x_2) \gg$ variance(σ_η)} for global mean temperature.

If $|\kappa| \ll 1$ or if $|\kappa| \gg 1$ the transfer function would essentially be one-dimensional and any of the updating rules would result in similar posterior probabilities of ruin. Hence we choose the non-trivial case $|\kappa| \approx 1$. The quantity σ_η represents another degree of freedom. As this article deals with the representation of imprecise prior knowledge and its updating and not so much with the uncertainty contained in the likelihood, the variance of the prior marginals should be much higher than the variance of y, and so we choose $\sigma_\eta \ll \sigma$, in particular, $\sigma_\eta = \sigma/10$ when we have to specify it. This is also in accordance with the experts' statements on CLIMBER-2 as mentioned in the paragraph above. Furthermore, as will become apparent below, for small σ_η, $|\kappa| = 1$ reveals a degenerate exception (compare Eq. (16)) for which reason we avoid such a choice. Whenever we do not display the dependency of results on κ but have to fix its value (e.g., for numerical results) we choose $\kappa := 1.05$.

As recalled in the Appendix A.1 a two-dimensional Gaussian prior P satisfies the constraints set by the marginals $\sim N(\mu, \sigma^2)$, iff there exists $f \in [-1, 1]$ with

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & f \\ f & 1 \end{pmatrix}, \tag{11}$$

[10] The more general case is easily obtained by replacing the Lebesgue measure with an appropriate dominating measure. In particular, the basic development also applies to the discrete case, where the corresponding integrals become simple sums. Note further in this context that if $P_i(\cdot)$, $i = 1,2$, is dominated by some σ-finite measure $\lambda_i(\cdot)$, $i = 1,2$, then every element of $\mathscr{P}$ is dominated by $\lambda_1(\cdot) \times \lambda_2(\cdot)$. In particular, since $P_1(\cdot)$ and $P_2(\cdot)$ are assumed to be absolutely continuous throughout the paper, also every $P \in \mathscr{P}$ shares this property and can be described by an appropriate Lebesgue density.

[11] For a multivariate application, the first entry would represent a vector of means, the second the symmetric covariance matrix.

[12] Note that the class is not convex.

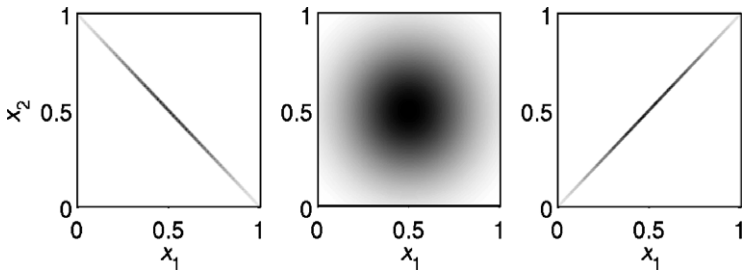

Fig. 1. Three extreme representatives of the class of Gaussian priors with prescribed marginals. From left to right: maximally anticorrelated case ($f = -1$), uncorrelated case ($f = 0$), and maximally correlated case ($f = 1$)– for a definition of the parameter f see Eq. (11). The maximally (anti)correlated cases are degenerate in the sense that the supports of the distributions are one-dimensional. This will lead to paradoxical inferences during Bayesian updating.

and $P \sim N((\mu,\mu)^t, \Sigma)$, whereby Σ denotes the covariance matrix of P and the results for $f = \pm 1$ are to be understood as the respective limit being taken.

Hence, we conveniently parameterize the class $\mathscr{P}$ of bivariate Gaussian priors with prescribed marginals by one single parameter $f \in [-1,1]$, which is simply the correlation coefficient of x_1 and x_2. Setting $f = 0$ represents the uncorrelated (standard), $f = \pm 1$ the maximally (anti)correlated case (see again Fig. 1).[13] In the following, we will call this $\mathscr{P}$ *correlation class*.

4.2. Posterior properties

With the specification of the likelihood (see Eq. (10)) and the parameterization of the correlation class, we can calculate the set of posterior distributions as a function of the correlation coefficient f and the observation y. To obtain the posterior marginal $P_{\text{post}}(x_1)$ for our quantity of interest x_1, we have to integrate the bivariate posterior over x_2, revealing (see Appendix A.2)

$$\begin{aligned} &P_{\text{post}}(x_1 \mid y) \sim N(\mu'(f,y), \sigma'^2(f,y)) \text{ with} \\ &\mu'(f,y) = \frac{\mu(1-(1-f)(\kappa-1)\sigma^2/\sigma_\eta^2) + (f+\kappa)y\sigma^2/\sigma_\eta^2}{1+(1+2f\kappa+\kappa^2)\sigma^2/\sigma_\eta^2}, \\ &\sigma'(f,y) = \sigma\sqrt{\frac{1+(1-f^2)\sigma^2/\sigma_\eta^2}{1+(1+2f\kappa+\kappa^2)\sigma^2/\sigma_\eta^2}}. \end{aligned} \tag{12}$$

We utilize this expression to calculate the posterior probability of ruin with respect to x_1

$$P^*_{\text{post},y}(f) = \int_{x_1^*}^{\infty} N(\mu'(f,y), (\sigma'(f,y))^2)(x_1)\mathrm{d}x_1. \tag{13}$$

The dependency of the probability of ruin on y is depicted in Fig. 2, upper graph, dotted curved line, for the assumption of independent parameters (uncorrelated case $f = 0$, i.e., $P(x_1,x_2) = P_1(x_1) \cdot P_2(x_2)$) and particular choices of $\kappa = 1.05$, threshold value $x_1^* = 0.95$, (beyond which ruin occurs) and variance of the observation $\sigma_\eta = \sigma/10$.[14]

The case of dominating likelihood uncertainty – not to be considered further – is obtained by $\sigma_\eta \to \infty$:

$$\begin{aligned} &\lim_{\sigma_\eta \to \infty} \mu' = \mu, \\ &\lim_{\sigma_\eta \to \infty} \sigma' = \sigma, \end{aligned} \tag{14}$$

i.e., if the measurement y becomes non-informative on $\kappa x_1 + x_2$, then the marginal prior and posterior on x_i are identical. We now apply the different learning rules in our setting.

4.3. Application of generalized Bayes' rule (GBR)

In order to derive the upper and lower probability of ruin according to GBR, we simply have to ask for the supremum and the infimum of $P^*_{\text{post},y}(f)$ over the correlation coefficient $f \in [-1,1]$ for given y. (While in [32] an algorithm is introduced for GBR on prior classes with prescribed marginals, we have to point out that the algorithm cannot be utilized here as our class is not convex.) We display the result as function of y in Fig. 2, upper graph, dashed-dotted lines. Both curves derived from GBR display quasi step-function type behavior.

[13] Fréchet [23] derived bounds for classes of priors with prescribed precise marginals. For our class it follows: Let F the cumulative marginal distribution, $f \in [-1,1]$, and H_f the cumulative distribution of the prior characterized by f. Then for any $x_1,x_2, \max(F(x_1) + F(x_2) - 1,0) \leqslant H_f(x_1,x_2) \leqslant \min(F(x_1),F(x_2))$ [23]. When observing that for all $x_1, f = \pm 1$, $H_f(x_1 + 1/2, \pm x_1 + 1/2) = F(x_1)$, we readily see that $f = -1, +1$ represent the lower and upper Fréchet bounds, respectively.

[14] As mentioned above, we will stick to these parameter values whenever values need to be specified in the remainder of this article. As to be expected, the probability of ruin increases monotonously with y.

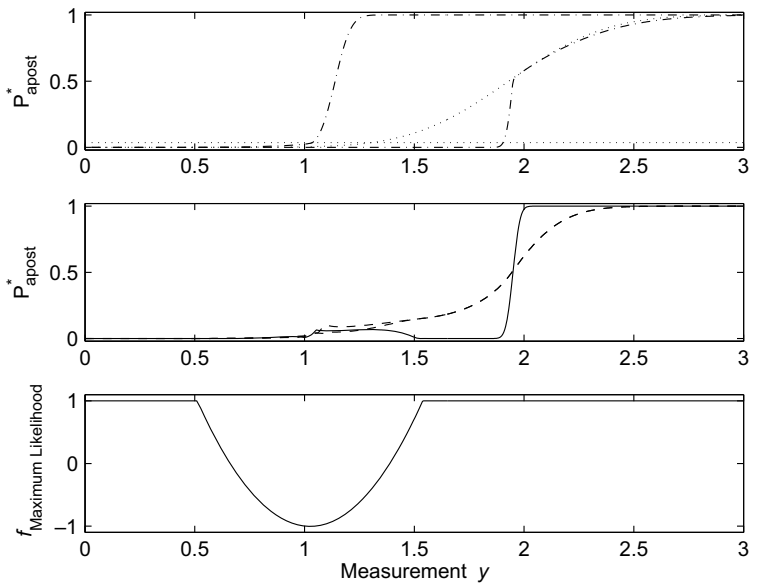

Fig. 2. Probability of ruin for the correlation class parameterized by the coefficient f (see Eq. (11)) for $\kappa = 1.05,\ x_1^* = 0.95$, $\sigma_\eta = \sigma/10$. *Top*: horizontal dotted line: prior value, curved dotted: posterior value for the (standard) case of independent parameters ($f = 0$), dashed-dotted: posterior bounds from the GBR that displays a quasi step-function-like behavior. *Center*: solid line: maximum likelihood estimate, dashed lines: weighted maximum likelihood estimate. The maximum likelihood estimate leads to rather low probabilities of ruin. The fact that the solid curve shows a non-monotonic relation between measurement and vastly deviates from its weighted counterpart (dashed) undermines trust in that updating method. *Bottom*: Correlation parameter f obtained from maximum likelihood update method, given y. For large y, the maximum likelihood update method prefers $f = 1$. Then Bayesian learning implies the intersection of the support of the likelihood with the line $x_1 = x_2$ (fully correlated), hence, the posterior concentrates all weight in a small line segment around a central point. Therefore, the probability of ruin must show a sharp transition (center graph, solid line) when this point crosses x_1^* as a function of y.

Apparently, GBR reveals much less informative results than the standard method would proclaim. In particular for $y \in [1.3, 1.8]$ GBR reveals *no information at all* on the posterior probability of ruin, i.e., we only know $P^*_{\text{post},y} \in [0, 1]$. Hence, in the GBR paradigm, utilizing the more realistic class of priors instead of the uncorrelated prior only, reveals drastically different results. The question is: how would the results change when less conservative (compared to the GBR) methods of updating are being used? Before we discuss these methods we would like to highlight the underlying reason for the non-informative features of GBR.

4.4. The illustrative limit $\sigma_\eta \to 0$

In Fig. 2, we have displayed results for the choice $\sigma_\eta = \sigma/10$, hence the variance of the likelihood function $P(y|x_1, x_2)$ is much smaller than the variance of the prior. We can expect to find the analytically transparent case $\sigma_\eta \to 0$ an illustrative limit for understanding the behavior of the posterior under GBR. If $\sigma_\eta \to 0$, then the support of $L(x_1, x_2) = P(y|x_1, x_2)$ collapses to the one-dimensional linear manifold that solves the equation $y = \kappa\, x_1 + x_2$. Furthermore,

$$\lim_{\sigma_\eta \to 0} \mu'(f, y) = \frac{\mu(1-f)(1-\kappa) + y(f+\kappa)}{1 + 2f\kappa + \kappa^2},$$
$$\lim_{\sigma_\eta \to 0} \sigma'(f, y) = \sigma\sqrt{\frac{1-f^2}{1 + 2f\kappa + \kappa^2}}. \tag{15}$$

Now consider the degenerate priors for $f = \pm 1$.

Eq. (15) imply $\lim_{|f| \to 1} \sigma' = 0$. In that limit, the prior P and the likelihood L_y represent two 1D lines in the 2D space spanned by x_1, x_2, intersecting only at one point, leaving no space for posterior uncertainty in x_1. Hence, the support of $P_{\text{apost}}(x_1)$ collapses to μ'. Therefore it is worthwhile to explicitly report μ' for these two extreme cases:

$$\mu'(f = -1, y) = \frac{y - 2\mu}{\kappa - 1}, \tag{16}$$
$$\mu'(f = +1, y) = \frac{y}{\kappa + 1}, \tag{17}$$

which can also be interpreted as the intersection of the lines $y = \kappa x_1 + x_2$ either with $1 = x_1 + x_2$ (see Fig. 3, for $f = -1$), or with $x_1 = x_2$ (for $f = 1$), i.e., the intersection of δ-type likelihood and (anti)correlated prior, respectively. From Eqs. (16) and (17) we conclude further (compare also Fig. 4):

$$P^*_{\text{post}}(f=-1) = \left\{ \begin{matrix} 0 & \text{for } y < (\kappa-1) & x_1^* + 2\mu \\ 1 & \text{for } y \geqslant (\kappa-1) & x_1^* + 2\mu \end{matrix} \right\}, \tag{18}$$

$$P^*_{\text{post}}(f=+1) = \left\{ \begin{matrix} 0 & \text{for } y < (\kappa+1) & x_1^* \\ 1 & \text{for } y \geqslant (\kappa+1) & x_1^* \end{matrix} \right\}. \tag{19}$$

These two equations explain the structural changes in $\overline{P}_{\text{post.GBR},y}$ (at $y \approx 1$) and $\underline{P}_{\text{post.GBR},y}$ (at $y \approx 2$), depicted as dashed-dotted lines in Fig. 2, upper graph (the positions of the discontinuities can easily be understood by noting $\kappa, x_1^* \approx 1$). Hence GBR is non-informative over a large interval of observations y (in our example for $y \in [\approx 1, \approx 2]$), i.e., we lose all information on the

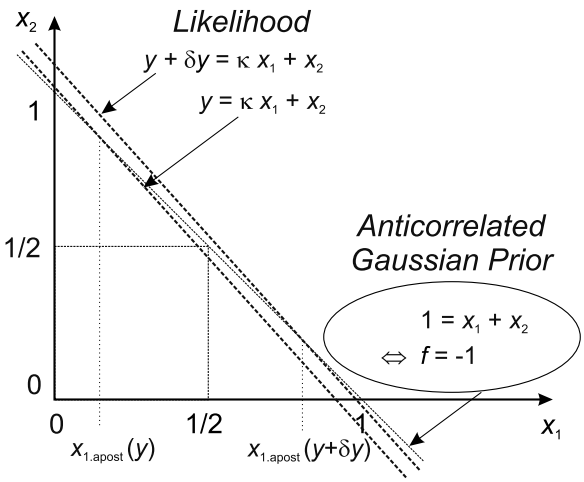

Fig. 3. Discussion of Bayesian learning for the double degenerate case $\sigma_\eta \to 0, f \to -1$ (fully anticorrelated prior). Bayesian learning is reduced to looking up the intersection of the lines "$1 = x_1 + x_2$" and "$y = \kappa x_1 + x_2$". Note that although the posterior uncertainty on x_1 is zero ("a well-defined intersection of lines"), the posterior value for x_1 strongly varies with mild variations in y. For $f = +1$, quite the contrary is the case. Hence, in many parameter settings, these two extreme cases tend to span a large interval for the probability of ruin for GBR, leading to non-informative results.

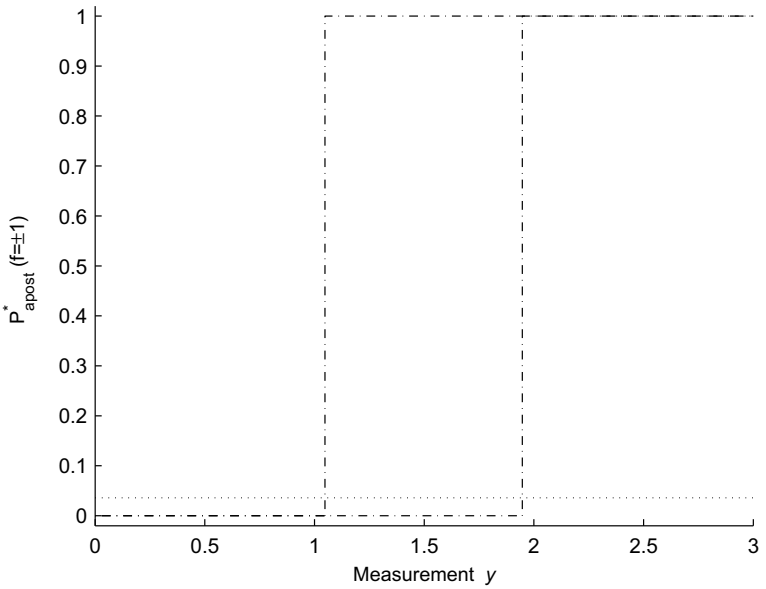

Fig. 4. Illustration of Eqs. (18) and (19): Bayesian learning for the two degenerate priors $P(f = \pm 1)$ and $\sigma_\eta \to 0$ (given the standard values $\kappa = 1.05, x_1^* = 0.95$). These two priors alone are sufficient to open the rather large non-informative window between $y \approx 1$ and $y \approx 2$ when GBR is used for updating. As before, the prior probability of ruin is indicated by the dotted line.

probability of ruin due to Bayesian updating of the correlation class. This is an extreme example of dilated posterior probability bounds (compared to the prior probability) over a significant range of observations (see Seidenfeld and Wasserman [40] for a discussion of the dilation phenomenon where posterior bounds are dilated for all possible measurements y). It should be noted that dilated posterior bounds also occur in the more realistic case of non-vanishing variance of likelihood function (see Fig. 2). In fact, the only likelihood function for which there exists no observation y that dilates the posterior probability of ruin is the non-informative likelihood function with $\sigma_\eta \to \infty$. We find this behavior of the GBR dissatisfying in this setting, and will investigate whether it could be avoided by more informative learning rules.

The priors with $f \to -1$ display a further type of "instability" that is not exhibited by $f \to +1$: Let $\kappa = :1 + \varepsilon$. Then $\mu'(f = -1, y) = (y - 2\mu)/\varepsilon$, according to Eq. (16). This implies for $\kappa \approx 1$, $\varepsilon \ll 1$ that Bayesian learning in the strongly anticorrelated limit is very unstable with respect to the measurement y even if only this single prior is considered.

For all these reasons, we will restrict the updating problem by (i) restricting the gradient of prior densities which would exclude $|f| \to 1$ (see Section 4.6), and (ii) implementing the more informative maximum likelihood upate rule and the new weighted likelihood upate rule for the unrestricted class of bivariate Gaussians with prescribed marginals.

4.5. Likelihood update results

4.5.1. Maximum likelihood updating

In order to calculate the results from the maximum update likelihood rule, we need to assess the prior density of observing y for a given prior $P_f(x_1, x_2) = N((\mu, \mu), \Sigma(f))(x_1, x_2)$ across the class of Gaussian priors parameterized by correlation $f \in [-1, 1]$. It is shown in the Appendix that the resulting density is again Gaussian with $P_f(y) = N(\mu_y, \sigma_y^2(f)), \mu_y = \mu(1 + \kappa)$, and $\sigma_y^2(f) = \sigma^2(1 + 2\kappa f + \kappa^2) + \sigma_\eta^2$. Note that the mean of $P_f(y)$ is independent of f while its variance is not, and it assumes values between $(\kappa \pm 1)^2\sigma^2 + \sigma_\eta^2$.

The maximum likelihood rule requires us to select only those priors which yield the highest weight $W_y(f) = P_f(y)$ on observing the particular y that actually occurs. By standard curve discussion we can show that $P_f(y)$ is maximized w.r.t. f by solving the equation $\sigma_y(f) = |y - \mu_y|$ in case there exists a solution over $f \in [-1, 1]$ or by maximization over $f \in \{-1, 1\}$ otherwise. Most importantly, the maximum is unique. Hence there exists a one-to-one mapping from observation y to correlation parameter f characterizing the prior $P_f(x_1, x_2)$ that maximizes the prior probability of y. Accordingly the maximum likelihood update rule selects only one prior which then is to be updated by means of Bayes' rule.[15] Consequently, the class of posteriors consists of exactly one element, for which reason we can drop the upper and lower bar for the corresponding probability of ruin $P^*_{\mathrm{ML},y}$, which is uniquely determined by the correlation parameter f.

The dependence of f on y is shown in Fig. 2, lower graph. Extreme values of y are most easily generated by $f = 1$ while central values of y are more preferred by $f = -1$. Between these two cases there is a continuous transition.

Fig. 2, center graph, shows the probability of ruin that results from the maximum likelihood update method. It yields a non-monotonous functional relation $P^*_{\mathrm{ML}}(y)$ that may appear counter-intuitive at a first glance. However the non-monotonous behavior can be understood when one relates the center to the lower graph. When discussing only y-branches with $f_{\mathrm{ML}}(y) = 1$, $P^*_{\mathrm{ML},y}$ is monotone as it must be. However, in between, around $y \approx \kappa$, the updating rule prefers the anticorrelated prior which must lead to a sharp switch in $P^*_{\mathrm{ML},y}$ as $y \approx \kappa$ is crossed (see Eq. (16) and remark afterwards). Hence the interplay of changing y as well as $f_{\mathrm{ML}}(y)$ leads to a non-monotonous $P^*_{\mathrm{ML},y}$.

4.5.2. Weighted likelihood updating

We now compare the results from the maximum likelihood update rule to the results from its weighted counterpart (see also Appendix A.5). In order to calculate the set of posteriors under the weighted likelihood update rule, we need to identify the level sets $\mathscr{P}(w_y)$ of priors which yield weight $W(f) := P_f(y) = w_y$ for a given observation y. Conveniently it turns out that each level set consists at most of two priors (in case there are two, their f-values would bracket f_{ML}).

As a key result, in Fig. 2, center graph, it becomes apparent that maximum (solid line) and weighted (dashed lines) likelihood update qualitatively deviate for $y > 1.5$. Fig. 5 illuminates the underlying reason. Large values of y force the maximum likelihood rule to select $f = 1$ which comes with $P^* = 0$ for $y < 1.9$ (see Eq. (19)). However, the center graph of Fig. 5 reveals that if one allowed for f mildly smaller than 1, $P^* = 0$ is *structurally unstable*, hence, a weighting method must result in much larger values for P^*, as shown in the lower graph. The lower graph furthermore illustrates the "purifying" mechanism within any of the likelihood methods: those priors resulting in $P^* = 1$ due to dilation-type behavior come with zero weight, hence their influence on P^* is eliminated by both likelihood methods (but not by standard GBR).

The fact that the maximum likelihood update drastically deviates from its weighted counterpart casts doubt on the reliability of the maximum likelihood update rule and gives the impression that it may be fundamentally "non-robust". One may argue that the non-monotone behavior may disappear once the class of priors is chosen more adequately – by avoiding extremely degenerate cases like $f = \pm 1$ that come along with diverging gradients. In the following Subsections we will see, however, that this is not the case; quite the contrary, any effect observed so far can be found again (although in a somewhat smoothed version) when gradients become restricted.

[15] So, in this context the type II maximum likelihood procedure and the maximum likelihood updating coincide.

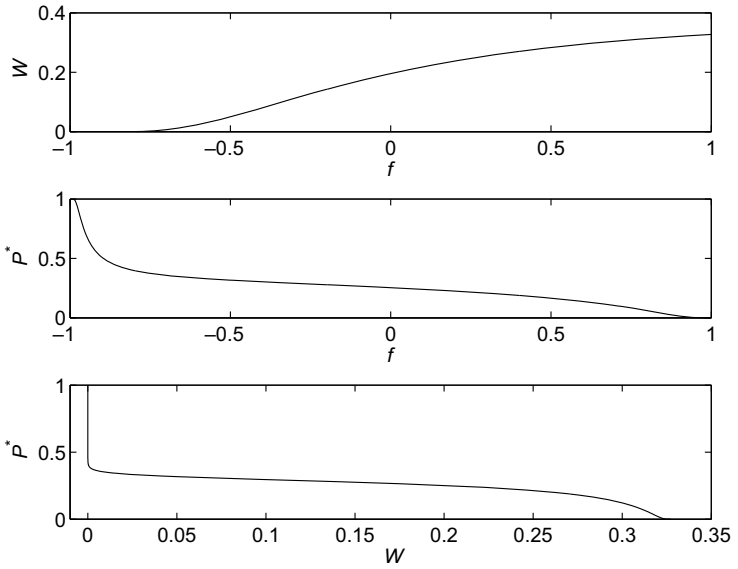

Fig. 5. Relation of weight function $W(.)$, parameter f specifying the prior, and probability of ruin, for the special case $y = 1.7, \kappa = 1.05, x_1^* = 0.95, \sigma_\eta = \sigma/10$. The maximum likelihood update rule requires us to select the $f = f_{ML}$, i.e., the prior for which W, the prior probability of y is largest. Hence, $f_{ML} = 1$ (see upper graph). This was to be expected as for the rather "large" value of y, the prior with highest correlation (i.e., $f = 1$) prefers the "extreme" y the most among all priors. However, $P^*(y = 1.7 f = 1) = 0$ (see Eq. (19) and center graph). The important point is that the case $f = 1$ is exceptional within the class of priors as for $f \in [-1,0.5]$, $P^* > 0.1$ (center graph). The bottom graph shows that when averaging $P^*(f(W))$ over the W-scale weighted with W, an average somewhere between 0.1 and 0.3 is to be expected (weighted likelihood update method), drastically differing from the maximum likelihood update.

4.6. Imposing constraints on gradients

Imposing constraints on the gradients basically results in a reduction of the class of Gaussians with prescribed marginals that is reflected in a reduction of the range for the correlation parameter f at its outer ends. The way how to infer from the expert's "resolution" parameter N on the admissible range of f is described in Appendix A.6.2.

Figs. 6 and 7 display the most extreme members of $\mathscr{P}$ if we set in (6) $N_1 := 5$ and $\beta := 0$ or =1, respectively (i.e., $N_2 := N_1$ or $N_2 := N_1^2$, respectively). In the latter case, we allow for "more" prior, mutually distinct opinions.

Figs. 8 (for $\beta = 0$) and 9 (for $\beta = 1$) reveal the effects of imposing gradient constraints on the class of priors. The curves for the probability of ruin $P^*(y)$ become smoother, in particular in the case $\beta = 0$, and more similar. However, still drastic differences between various updating methods remain:

- Standard Bayesian learning (i.e., the uncorrelated case, upper graph, curved dotted line) results in a more optimistic estimate than given by the upper probability of ruin from GBR.
- For $1.4 < y < 1.8$, the upper probability of ruin according to the weighted likelihood method (lower graphs, dashed curves) exceeds the estimate according to the maximum likelihood update method (lower graph, solid line), in part by an order of magnitude. It demonstrates that maximum likelihood update for the class of Gaussian priors is *not* a structurally robust learning rule (in line with the discussion in Fig. 5). This finding casts doubt on results obtained by that method and we do not recommend adopting the maximum likelihood update rule.

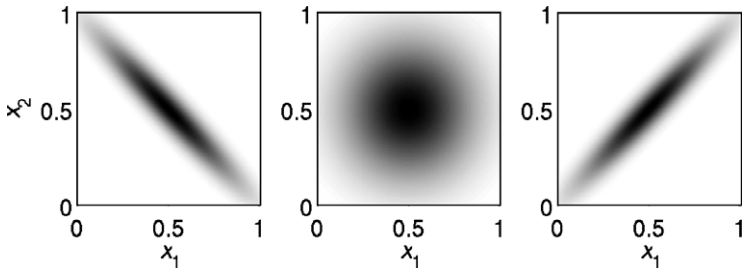

Fig. 6. Extreme members of the class of priors for a bounded gradient condition, consistent with 5 blocks in the 2D parameter space ($N_2 = N_1 = 5 \Rightarrow \beta = 0$ along the notation of Eq. (6). Left graph for minimum f, center for the (standard) uncorrelated case ($f = 0$), right graph for maximum f.

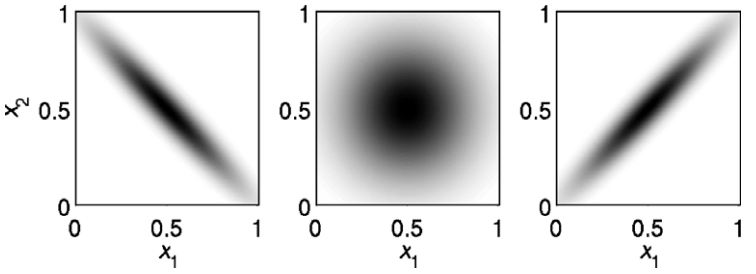

Fig. 7. The same as in Fig. 6, yet for dimension-adjusted resolution (i.e., $\beta = 1$): $N_2 := N_1^2 = 5^2$. Note that in higher dimensions n (here: $n = 2$) the prescriptions for N_n according to the present versus the previous graph would diverge the more, the larger n. We propose that a realistic description prescriptions for N_n would imply a compromise between these extremes of spatial resolution that reflect the degree of sophistication expert options may display.

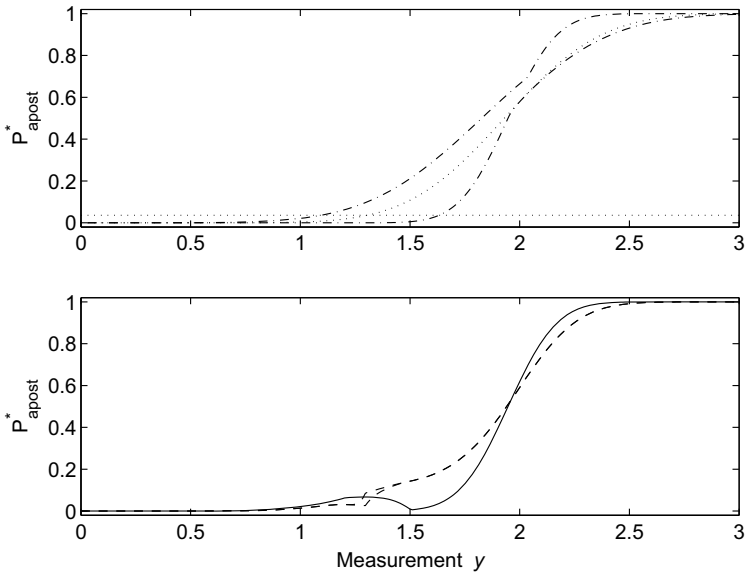

Fig. 8. The same as Fig. 2, however, for bounded gradients according to N_2: = 5 $\Leftrightarrow$ $\beta = 0$ in (6). Bounding of the gradients reveals much softer curves. Note, however, that even for this class of priors "regularized" by the gradient filter, maximum likelihood update strongly deviates from weighted likelihood update. This demonstrates how questionable it may be to use maximum likelihood update–that is based on a single prior in our case–*if* one desires a more balanced (through weighting) influence of all the priors.

4.7. Consequences of updating results for a stylized decision-maker

For the following interpretation one may imagine a stylized, rather conservative insurance company that offers contracts only clients with an upper probability of ruin below a given threshold. The latter depends on the number of clients to pool the risk with and the upper probability of the company's ruin, the company is willing to accept. For illustrative purposes, in the following we assume two cases of upper limits per client the company is willing to accept: $\overline{P}^* = 13\%$ or 27%.[16] The point is now that the company has no direct access to a client's $\overline{P}^*$ but needs to indirectly infer on $\overline{P}^*$ by observing the "client characteristic y," that could be "measured" in principle.

Fig. 10 highlights the link between the upper probability of ruin and the maximum admissible client characteristic y^*, for any of the learning rules discussed in this paper. We have indicated $\overline{P}^* = 27\%$ by a horizontal solid line. From the four intersecting curves $\overline{P}^*(y)$ we can then deduce the maximum y, i.e., y^*, for which the insurance company would agree to insure a client (for the intermediate gradient filter $\beta = 1$). The y^* values for $\overline{P}^* = 13\%$ or 27% and the four learning rules are summarized in Fig. 11 and Table 1.

Obviously, y^* is a strong function of the learning rule. The ordering is in part also a function of the admissible P^*. GBR results in the most exclusive, maximum likelihood in the most inclusive insurance policy. Clients with $1.37 < y < 1.90$ would be rejected due to GBR, but insured according to the maximum likelihood rule. Of the four learning rules, GBR is the one that

[16] An illustration of those numbers can be found in [28, Section 5].

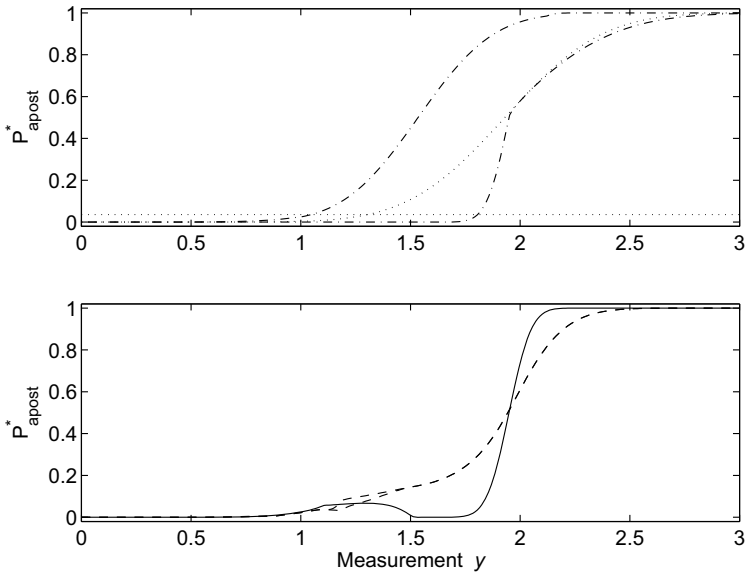

Fig. 9. The same as the previous figure, yet for a larger number of independent expert opinions, i.e., $N_2 := 5^2 \Leftrightarrow \beta = 1$. The curves describe an intermediate case between Figs. 2 and 8.

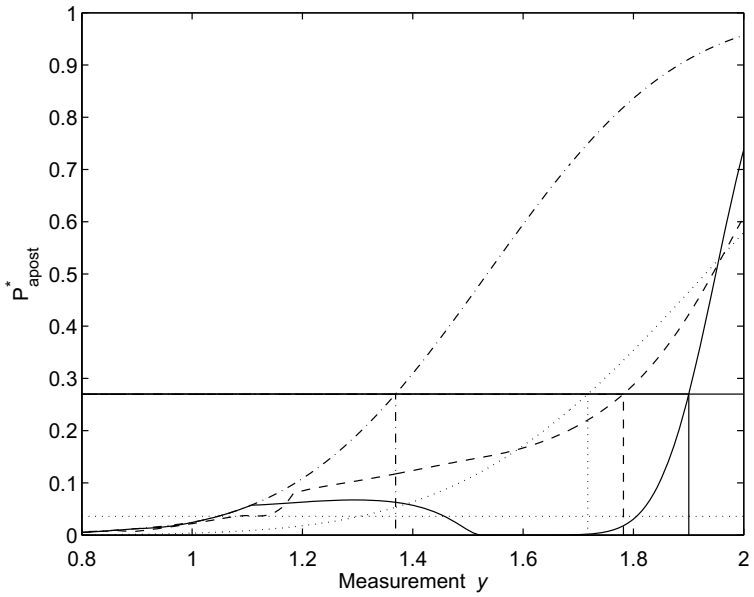

Fig. 10. Decision implications of the results from Fig. 9 to be assessed by a stylized insurance company requesting an upper probability of ruin of 27% (horizontal solid line). The line styles are again as follows. Dashed-dotted: GBR; dashed: (new) weighted likelihood; dotted: standard precise method (based on assuming independent marginals); solid: maximum likelihood update rule. For any of the four learning rules, we indicate the maximum characteristic y^* with which a client would be insured by a vertical line. Obviously, y^* is a strong function of the learning rule. GBR would lead to the most exclusive insurance practice. When the company accepted the weighted likelihood update method (dashed line), clients with $1.37 < y < 1.78$ would be insured in addition.

certainly does not add spurious information. If the insurance company found GBR not informative enough, however, as there might be too few clients with $y \leqslant y^*_{\mathrm{GBR}}$, out of the remaining three learning rules we would recommend adoption of our new weighted likelihood update rule as the most robust variant. Its superiority compared to the maximum likelihood update rule has already been discussed above. Compared to the standard Bayesian updating, assuming independent marginals, large differences in the assessment of the probability of ruin exist particularly for $y < 1.2$, i.e., the regime of small probabilities. There the standard method estimates the probability of ruin to be up to an order of magnitude lower than indicated by the weighted likelihood update rule.

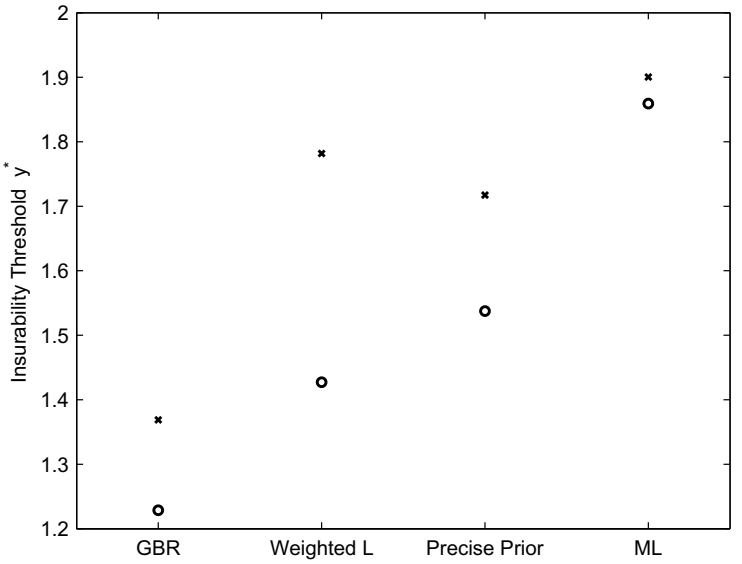

Fig. 11. y^*, the upper limit of characteristics y with which clients would be insured in dependence on different learning rules und number of clients: Circles: $\overline{P}^* = 13\%$; crosses: $\overline{P}^* = 27\%$. The abscissa indicates the four learning rules considered in this paper. Weighted L: (new) weighted likelihood update rule; Precise Prior: standard Bayesian updating starting with a precise uncorrelated prior; ML: maximum likelihood update rule. It is remarkable that–for $\overline{P}^* = 27\%$– more than 50% of the gap between the robust, yet least informative GBR and the most optimistic, yet non-robust maximum likelihood update rule can be regained when using the new weighted likelihood update rule. (As for previous Figures all entries are for $\kappa = 1.05, x_1^\dagger = 0.95, \sigma_\eta = \sigma/10$).

Table 1
y^*, the upper limit of "measurements" y with which clients would be insured in dependence on different learning rules and admissible probabilities of ruin (per client)

	$\overline{P}^*$	13%	27%
	Updating rule	y^*	y^*
1	Generalized Bayes rule (GBR)	1.23	1.37
2	Weighted likelihood	1.43	1.78
3	Uncorrelated prior	1.54	1.72
4	Maximum likelihood	1.86	1.90

In summary, the new learning rule may be a convincing method through which the company could insure additional clients that were rejected by GBR.

5. Summary and discussion

Using a stylized Gaussian analysis, this article addresses the chain of prevailing conceptual challenges one faces when attempting to update multivariate prior knowledge on adjustable deterministic models: the chain made-up by the formal representation of imprecise prior knowledge over ex-ante re-adjustment of overly conservative priors to some sort of generalized Bayesian updating of that prior.

We have set up a model for subjective uncertainty that aims at reflecting the opinions held by many Earth system modelers concerned with climate, biosphere and economic systems. By means of a questionnaire we found that the experts express much more confidence in marginals of priors (on model parameters) rather than in the correlation structure of those priors. Based on this insight, we have studied an imprecise prior defined by prescribing the marginals only. We have presented a show-case for Bayesian updating on the basis of that imprecise prior, based on the simplest multi-dimensional transfer function possible: $y = \kappa x_1 + x_2$ in combination with Gaussian probability density functions. To accomplish Bayesian learning we utilized two generalizations of Bayes' rule to imprecise priors that have appeared in the literature and also introduced a new method.

In particular we have considered the following updating rules which – even in our rather elementary example – unfolded rich posterior properties:

(1) Updating along the lines of generalized Bayes' rule (GBR), under which each member of the prior class is updated, and then the extremes among the posteriors are selected to calculate the lower and upper bound on the quantity of interest.

(2) The maximum likelihood update method, which focuses on those priors that maximize the prior probability of observing the actual measurement y. It leads to more informative results than GBR. However, we find the rule hard to justify as it completely disregards even those priors which may perform in the prediction only infinitesimally less well than the optimal priors. Also this leads to counter-intuitively large posterior sensitivity as well as large non-monotonicity with respect to the observational evidence.
(3) For that reason we have introduced a weighted likelihood update method that considers all priors, yet weights their influence on the posterior result. When applied to the class of priors that appears as most realistic in practice (i.e., the class of gradient-limited Gaussians) we find that the two likelihood methods may deviate by an order of magnitude in their estimate of the posterior probability of ruin.
The weighted likelihood method carries the drawback (or advantage) that the weighting function implies a degree of freedom that is subject to normative decisions. We have chosen a weighting proportional to the prior probability of the measurement. Both likelihood methods are more informative than GBR; however, they share the disadvantage that they may add spurious information if the class of priors is overly inclusive (i.e., contains incompetent expert opinions), in contrast to GBR.
(4) For comparison we also considered the naive Bayesian solution based on the single prior derived from simply assuming independence. This typically leads to much lower probabilities of ruin, when compared to GBR or the weighted likelihood method.

Here we propose to consider GBR as the most conservative and easiest to justify rule first: if the class of priors is contaminated with unjustified elements, GBR nevertheless does not produce spurious information due to those elements. If all elements can be justified, the more informative weighted likelihood update method may be used, although it is harder to interpret.

In case one distrusts some of the elements contained in the set of priors used for calculation (so that both likelihood methods may produce spurious information) and GBR turns out non-informative, one may try to eliminate those priors. A subtle as well as practically relevant point here is that in fact we have reasons to believe that degenerate priors (in our example strongly (anti-)correlated priors) most likely are not members of the class that would perfectly model prior knowledge of an expert: usually there is a level of sophistication one thinks experts may be capable of in terms of resolution in parameter space. We removed degenerate cases by imposing an upper bound for the gradient of the priors. Thereby, we obtained more informative results for GBR and also felt justified to utilize maximum likelihood methods. However, those major "filtering efforts" on the class of priors do not address the very fact that in applications, it will never be possible to *perfectly* model an expert's knowledge, even after filtering. Along the lines of robust statistics one may then want to study posterior robustness with respect to miniscule changes in the prior. In the present article, we suppress this additional layer of complexity in order to distil the pure effects of changing updates rules.

Our new updating rule showed quite attractive behavior in the setting considered here: it is informative, more robust against small changes in observations and less in conflict with monotonicity (posterior vs. observation) than maximum likelihood updating. A skeptic may argue that any updating rule which discounts precise priors in view of the evidence before applying GBR would be logically inconsistent, as the evidence were used twice: firstly, the evidence is used to discount priors, secondly, those ranked priors are then updated with the evidence, and aggregated. This counter-argument would apply to Gilboa's and Schmeidler's rule as well as for our new rule. A certainly desirable axiomatic analysis is beyond the scope of this rather illustrative, yet conceptual paper. However, we observe that society very often behaves just like that on a "group-of-experts-level": it would listen more carefully to experts (i.e., precise priors) who have predicted the evidence with greater skill. Hence we observe that there exist normative settings in society that operate just along lines very similar to that rule. Furthermore, Gilboa's and Schmeidler's rule closely resembles a type II maximum likelihood prior, well-established within robust Bayesian statistics [4, p. 99]. In any case we find this type of using the evidence twice (which is not be confused with falsely applying Bayes rule twice to the same data) more convincing than disregarding the statements of experts – that they have no clue on correlations – and assuming an almost precise prior model for updating under GBR, as suggested in [34, Section 3.3 and references therein].

Of course, more detailed studies are needed to evaluate the rule in a more comprehensive manner. For such future research it will be important to investigate in particular (1) how the new updating rule will perform under the curse of dimension, i.e., for increasing number of parameters, (2) how robust its posteriors are with respect to miniscule changes in the prior, (3) what posteriors it will produce when applied to a nonparametric, presumably convex class, and (4) how one could transfer our innovations to complex dynamical models such as climate models.

Acknowledgements

First of all, we would like to thank half a dozen modelers from the fields of climate science, ecological modeling and economic growth theory, who volunteered to take part in our survey on prior beliefs of uncertainty model parameters: E. Bauer, N. Bauer, A. Ganopolski, D. Gerten, M. Hofmann, K. Lessmann, S. Schaphoff. Furthermore we would like to thank R. Klein and J. Schellnhuber for drawing our attention to the issue of purely known correlations in complex systems. Finally, H.H. and E.K. gratefully acknowledge support by the Volkswagen Foundation under grant number II/78470, T.A. by the Deutsche Forschungsgemeinschaft within the SFB 386. We are indebted to the Area Editor and two anonymous referees for their very helpful comments addressing several important issues, and to Teddy Seidenfeld for stimulating discussions.

Appendix A. Analytic treatment for Gaussian priors

There is numerous analytic work on imprecise Gaussians – as examples we just mention [4,14]. However this Appendix shall provide a self-contained set of equations that would allow the reader to verify the results in the main text straightforwardly.

A.1. Parameterizing the class of priors

First we recall a well-known relationship on how to infer from joint on marginal properties (see e.g., [8, p. 283]):

Lemma 1. *Let, for some $p_1, p_2 \in \mathbb{N}$, the variable $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} =: (x_1^t, x_2^t)$ with $x_1 \in \mathbb{R}^{p_1}$ and $x_2 \in \mathbb{R}^{p_2}$ distributed according to a multivariate Gaussian distribution with mean $\mu = (\mu_1, \mu_2)^t$ and covariance matrix $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ abbreviated as $x \sim N(\mu, \sigma)$ with $\mu_1 \in \mathbb{R}^{p_1}$ and $\Sigma_{11} \in \mathbb{R}^{p_1} \times \mathbb{R}^{p_1}$ then a the marginal density of x_1 is a (multivariate) Gaussian distribution with mean μ_1 and covariance matrix Σ_{11}, i.e., $x_1 \sim N(\mu_1, \Sigma_{11})$.*

In the following we will use the compact notation $(\cdot|\cdot)$ for scalar products and $(|M|\cdot)$ for quadratic forms w.r.t. a symmetric matrix M, i.e., $(|M|v) = v^t M v$.

Let P a two-dimensional Gaussian from our correlation class with the marginals $P_1 \equiv P_2 \sim N(\mu, \sigma^2)$. Then it can be expressed as

$$P(x) = c\mathrm{e}^{-\frac{1}{2}(|\Sigma^{-1}|(x-\bar{x}))}, \quad c = \frac{1}{(2\pi)\sqrt{\det \Sigma}}, \tag{A.1}$$

$\bar{x}$ denoting the mean, Σ the covariance matrix (see, e.g., [4]).

From Lemma 1, we immediately conclude $(\mu, \mu)^t = \bar{x}$ and $\Sigma_{11} = \Sigma_{22} = \sigma^2$. Furthermore we note that Σ must be symmetric as a covariance matrix. Hence, it remains to show that $f \in [-1, 1]$.

In order to do so, we exploit the fact that a symmetric matrix Σ defines a Gaussian in Eq. (A.1) iff it is positive semi-definite (as a covariance matrix must be), hence its two eigenvalues $\lambda_{1,2} \geqslant 0$. By deriving $\lambda_{1,2} = \sigma \cdot (1 \pm |f|)$, we conclude

$$\{\Sigma \text{ positive semi-definite}\} \iff \{|f| < 1\}. \tag{A.2}$$

A.2. Derivation of the posterior

A.2.1. Derivation of the bivariate posterior

Let $h_y := 1/\sigma_\eta^2$ (i.e., the "precision" of the likelihood), $x := (x_1, x_2)^t, \bar{x} = (\mu, \mu)^t, k := (\kappa, 1)^t$. Then according to Bayes' rule, with sorting quadratic and linear terms in x

$$P_{\text{post}}(x_1, x_2) \propto N((\mu, \mu), \Sigma)(x_1, x_2) \cdot N(y, 1/h_y)(\kappa x_1 + x_2) \tag{A.3}$$

$$\propto \mathrm{e}^{-\frac{1}{2}Q'} \text{ with} \tag{A.4}$$

$$Q' := (|A|x) - 2(\gamma|x) \quad \text{and} \tag{A.5}$$

$$A := \Sigma^{-1} + h_y k \otimes k^t, \tag{A.6}$$

$$\gamma := \Sigma^{-1}\bar{x} + y h_y k. \tag{A.7}$$

Since

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & f \\ f & 1 \end{pmatrix}, \text{ it follows that} \tag{A.8}$$

$$\Sigma^{-1} = \Gamma \begin{pmatrix} 1 & -f \\ -f & 1 \end{pmatrix} \text{ with } \Gamma := \frac{1}{\sigma^2(1-f^2)}. \tag{A.9}$$

In order to transform Q' to standard form we transform the x-coordinates according to

$$Q' = (|A|(x - x_0)) + C), \tag{A.10}$$

where x_0 and C do not depend on x. We determine x_0 by differentiating Eq. (A.10) w.r.t. x and obtain

$$x_0 = A^{-1}\gamma, \quad \text{hence} \tag{A.11}$$

$$P_{\text{post}}(x_1, x_2) = N(A^{-1}\gamma, A^{-1})(x_1, x_2). \tag{A.12}$$

A.2.2. Derivation of the posterior marginal in x_1

From the bivariate posterior in Eq. (A.12) we easily obtain the marginal in x_1 when we recall Lemma 1. Eq. (12) are then obtained through Eq. (A.12) and the previous definitions by symbolic manipulation in MATHEMATICA 5.2.

A.3. Derivation of the prior probability density of y

Recall that the prior density on the parameters x_1, x_2 is defined by

$$P(x_1, x_2) = N((\mu, \mu), \Sigma)(x_1, x_2). \tag{A.13}$$

Let $x := (x_1, x_2)^t, \bar{x} = (\mu, \mu)^t, k := (\kappa, 1)^t$. Set $F := (k|x) = \kappa x_1 + x_2$. Utilizing the following Lemma (see, e.g., [44, pp. 22,40])

Lemma 2. *Let x, γ denote n-dimensional vectors, $P(x)$ a probability density function for x. Let $y := (\gamma|x)$ for all x. Then*

(1) $\text{mean}(y) = (\gamma|\text{mean}(x))$,
(2) $\text{covar}(y) = (|\text{covar}(x)|\gamma)$.

We then know

$$P_F(F) = N(\mu_F, \sigma_F^2)(F) \quad \text{with} \tag{A.14}$$
$$\mu_F = (k|(\mu, \mu)^t) = \mu(1 + \kappa), \tag{A.15}$$
$$\sigma_F = \sqrt{(k|\Sigma|k)} \tag{A.16}$$
$$= \sigma\sqrt{1 + 2\kappa f + \kappa^2}. \tag{A.17}$$

Then we recall that $P(y|x) = N(F(x), \sigma_\eta^2)(y)$, hence

$$P_y \sim N(\mu_F, \sigma_y^2) \text{ with } \sigma_y := \sqrt{\sigma_F^2 + \sigma_\eta^2}. \tag{A.18}$$

For subsection A.4 we will further need the following relation: by means of Eq. (A.17) we show readily that

$$\inf_{f\in[-1,1]} \sigma_y = \sqrt{(\sigma(\kappa - 1))^2 + \sigma_\eta^2} =: \sigma_{y.\min}, \tag{A.19}$$
$$\sup_{f\in[-1,1]} \sigma_y = \sqrt{(\sigma(\kappa + 1))^2 + \sigma_\eta^2} =: \sigma_{y.\max}. \tag{A.20}$$

A.4. Derivation of the maximum likelihood priors

$P_y(y) < \infty$ sets also the weight function $W(y)$ according to which we preselect priors for the maximum likelihood updating rule: $W \equiv P_y$. Let μ_y,σ_y be the mean and standard deviation of the prior distribution for y. The present class of priors is conveniently parameterized by f which influences σ_y but not μ_y. By standard curve discussion we find that for given y, $W(f)$ is maximized if

$$\sigma_{y.\text{ML}} = |y - \mu_y| \tag{A.21}$$

in case that equation has a solution for $f \in [-1, 1]$. If we conclude from σ_y on f, we obtain

$$f_{\text{ML}} = \left\{ \begin{array}{ll} -1 & \text{for } |y - \mu_y| < \sigma_{y.\min} \\ +1 & \text{for } |y - \mu_y| > \sigma_{y.\max} \\ \frac{(y-\mu_F)^2 - \sigma_\eta^2}{2\kappa(\sigma^2 - 1 - \kappa^2)} & \text{otherwise} \end{array} \right\}. \tag{A.22}$$

A.5. Derivation of the weighting likelihood result

We note that the derivative $W'(f)$ vanishes at the maximum once over $[-1, 1]$. namely for the term given in Eq. (A.22). Hence, the equation $W(f) = w$, for given w, can have at most one solution left, and one right from f_{ML}. Either solution is found numerically by specifying $[-1, f_{\text{ML}}]$ and $[f_{\text{ML}}, 1]$ as search intervals. Then we discretize the space for w between $[0, W(f_{\text{ML}})]$ and apply Eq. (5).

A.6. Derivation of the maximum-derivative condition

Let a Gaussian probability density function P be given as

$$P(x) = c\mathrm{e}^{-\frac{1}{2}(|\Sigma^{-1}|x)}, \quad c = \frac{1}{(2\pi)\sqrt{\det \Sigma}}, \tag{A.23}$$

x being a two dimensional vector (see [4]).

A.6.1. Derivation of the maximum gradient of a Gaussian density function

Let P be a bivariate Gaussian density function as defined in Eq. (A.1).

From elementary manipulations one establishes (by "Σ^{-2}" denoting the square of the inverse of Σ)

$$G(x) := |\mathrm{grad}P|^2(x) = P^2(x)(|\Sigma^{-2}|x). \tag{A.24}$$

The maximum of the function $G(x)$ will serve as criterion for whether P will be considered as a member of the prior class. In order to determine the maximum of G, we establish the necessary condition for a local maximum:

$$\forall_{i\in\{1,2\}} \quad \frac{\partial}{\partial x_i} G = 2P\frac{\partial}{\partial x_i} P(x|\Sigma^{-2}|x) + P^2 2(e_i|\Sigma^{-2}|x) = 0, \tag{A.25}$$

where e_i denotes the unit vectors of the coordinate system. Then we note that without loss of generality we can choose the unit vectors identical with the normalized eigenvectors v_i of Σ^1 (to which may also belong the eigenvalues λ_i). We conclude

$$\forall_{i\in\{1,2\}} \quad -\lambda_i(v_i|x)(x|\Sigma^{-2}|x) + \lambda_i^2(v_i|x) = 0. \tag{A.26}$$

Case 1. $\lambda_1 \neq \lambda_2$:

Without loss of generality we assume $\lambda_1 > \lambda_2$.

Case 1.1. $(v_1|x) \neq 0$:

Then we conclude

$$(x|\Sigma^{-2}|x) = \lambda_1. \tag{A.27}$$

Case 1.1.1. $(v_2|x) \neq 0$:

Then we conclude

$$(x|\Sigma^{-2}|x) = \lambda_2. \tag{A.28}$$

However, as $\lambda_1 \neq \lambda_2$, the last two equations cannot be fulfilled simultaneously, hence, Case 1.1.1 can be ruled out.

Case 1.1.2. $(v_2|x) = 0$:

In summary for Case 1 we can conclude that x must be parallel to one of the eigenvectors:

$$\exists_{i\in\{1,2\}} \quad x = \alpha_i v_i \quad \text{with} \quad \alpha_i = 1/\sqrt{\lambda_i}, \tag{A.29}$$

hence, the local maxima of G are along the eigenvectors at the standard deviations. It is then easily verified that the global maximum of G is along the larger eigenvector.

Case 2. $\lambda_1 = \lambda_2$:

The maximization problem can be reduced to a one-dimensional one due to rotational symmetry in x-space. By identifying the radial coordinate with α_1 of Case 1, one finds that Eq. (A.29) holds for Case 2 as well.

A.6.2. Implementation of the gradient constraint

We now use the above information in order to preselect members of the class of priors according to their maximum norm of the gradients of their densities.

As a reference, we use the maximum gradient of the marginals which in our example are both $\sim N(\mu, \sigma^2)$. Their maximum gradient G^* is readily derived as

$$G^* = \frac{1}{\sqrt{2\pi e}} \cdot \frac{1}{\sigma^2}. \tag{A.30}$$

Let p the dimension of x (in our example $p = 2$). Then for $p > 1$, the gradient of P will have a different unit than that of one of the marginals (in case x has a unit). Hence, a meaningful (in terms of units) restriction of the gradient reads as

$$|\mathrm{grad}P| \cdot (\Delta x)^{p-1} < N \cdot G^*, \tag{A.31}$$

where Δx denotes the typical scale per coordinate (in our example $\Delta x = 1 = 4\sigma$), N the "expert's resolution" (see Eq. (6)). The factor $(\Delta x)^{p-1}$ can be interpreted as follows: first, it adjusts units. Second, for a P whose density fills Δx in $p-1$ coordinates while being denser in a single coordinate, the above expression in essence reveals the 1D gradient along the eigenvector v^* with the largest eigenvalue λ^* of Σ^{-1}. If P were higher concentrated in further dimensions as well, the left-hand side would became larger. Hence, the above expression reveals the p-dimensional resolution, equivalent to N in Eq. (6).

In order to determine |grad$P$| at its maximum, we consider the 1D-function of α, $P((\mu,\mu)+\alpha v^*)$, if v^* is the eigenvector for the maximum eigenvalue λ^* of Σ^{-1}. Let $\sigma' := 1/\sqrt{\lambda^*}$. Then $P((\mu,\mu)+\alpha v^*) = c\sqrt{2\pi}\sigma' N(0,\sigma'^2)(\alpha)$, hence the maximum modulus of gradient of P reads $c\sqrt{2\pi}\sigma' \cdot 1/(\sqrt{2\pi e}\sigma'^2)$. Combining this information with Eq. (A.31) leads to a test on σ', and, in turn on Σ and on f.

References

[1] T. Augustin, On the suboptimality of robust Bayesian procedures from the decision theoretic point of view – a cautionary note on updating imprecise previsions, in: Bernard et al. [7], pp. 31–45.
[2] T. Augustin, F. Coolen, Nonparametric predictive inference and interval probability, Journal of Statistical Planning and Inference 124 (2004) 251–272.
[3] M.J. Bayarri, M.H. DeGroot, Optimal reporting of predictions, Journal of the American Statistical Association 84 (1989) 214–222.
[4] J.O. Berger, Statistical Decision Theory and Bayesian Analysis, Springer, 1985.
[5] J.O. Berger, D. Rios Insua, F. Ruggeri, Bayesian robustness, in: D. Rios Insua, F. Ruggeri (Eds.), Robust Bayesian Analysis, Lecture Notes in Statistics, vol. 125, Springer, New York, 2000, pp. 1–32.
[6] D. Berleant, Jianzhong Zhang, Using Pearson correlations to improve envelopes around the distribution of functions, Reliable Computing 10 (2004) 139–161.
[7] Jean-Marc Bernard, Teddy Seidenfeld, Marco Zaffalon (Eds.), ISIPTA'03, Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications, Lugano, Switzerland, July 14–17, 2003, Proceedings in Informatics, vol. 18, Carleton Scientific, 2003.
[8] G.E.P. Box, G.M. Jenkins, G.C. Reinsel, Time Series Analysis – Forecasting and Control, third ed., Prentice-Hall International, Inc., New Jersey, USA, 1994.
[9] R.M. Cooke, Experts in uncertainty, Opinion and Subjective Probability in Science, Oxford University Press, 1991.
[10] F.P.A. Coolen, Imprecise conjugate prior densities for the one-parameter exponential family of distributions, Statistics and Probability Letters 16 (1993) 337–342.
[11] F.P.A. Coolen, Statistical modeling of expert opinions using imprecise probabilities, PhD thesis, Eindhoven University of Technology, 1994.
[12] Fabio Gagliardi Cozman, Robert Nau, Teddy Seidenfeld (Eds.), ISIPTA'05, Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications, July 20–23, 2005, SIPTA, Carnegie Mellon University, Pittsburgh, PA, USA, 2005.
[13] W. Cramer, A. Bondeau, F.I. Woodward, I.C. Prentice, R.A. Betts, V. Brovkin, P.M. Cox, V. Fisher, J. Foley, A.D. Friend, C. Kucharik, M.R. Lomas, N. Ramankutty, S. Sitch, B. Smith, A. White, C. Young-Molling, Global response of terrestrial ecosystem structure and function to CO_2 and climate change: results from six dynamic global vegetation models, Global Change Biology 7 (2001) 357–373.
[14] A. DasGupta, W.J. Studden, Frequentist behaviour of robust bayes estimates of normal means, Statistics and Decision 7 (1989) 333–361.
[15] G. de Cooman, J. Vejnarova, M. Zaffalon (Eds.), ISIPTA'07: Proceedings of the Fifth International Symposium on Imprecise Probabilities and their Applications, Manno, CH. SIPTA, 2007.
[16] A.P. Dempster, A generalization of Bayesian inference, Journal of the Royal Statistical Society, Series B 30 (1968) 205–247.
[17] J. Dhaene, M. Denuit, M.J. Goovaerts, R. Kaas, D. Vyncke, The concept of comonotonicity in actuarial science and finance: theory, Insurance: Mathematics and Economics 31 (2002) 3–33.
[18] D. Dubois, H. Prade, Possibility Theory: An Approach to Computerized Processessing of Uncertainty, Plenum Press, New York, 1988.
[19] D. Dubois, H. Prade, Focusing versus updating in belief function theory, in: R.R. Yager, M. Fedrizzi, J. Kacprzyk (Eds.), Advances in the Dempster–Shafer Theory of Evidence, Wiley, New York, 1994, pp. 71–95.
[20] O. Edenhofer, N. Bauer, E. Kriegler, The impact of technological change on climate protection and welfare: insights from the model MIND, Environmental Economics 54 (2005) 277–292.
[21] C.E. Forest, P.H. Stone, A.P. Sokolov, M.R. Allen, M.D. Webster, Quantifying uncertainties in climate system properties with the use of recent climate observations, Science 295 (2002) 113–117.
[22] M.J. Frank, R.B. Nelsen, B. Schweizer, Best-possible bounds for the distribution of a sum – a problem of Kolmogorov, Probability Theory and Related Fields 74 (1987) 199–211.
[23] M. Fréchet, Sur la distance de deux lois de probabilité, vol. 6, Publ. Inst. Statist. Univ., Paris, 1957. pp. 185–198.
[24] A. Ganopolski, V. Petoukhov, S. Rahmstorf, V. Brovkin, M. Claussen, A. Eliseev, C. Kubatzki, CLIMBER-2: a climate system model of intermediate complexity. Part II: model sensitivity, Climate Dynamics 17 (2001) 735–751.
[25] P.H. Garthwaite, J.B. Kadane, A. O'Hagan, Statistical methods for eliciting probability distributions, Journal of the American Statistical Association 100 (2005) 680–700.
[26] I. Gilboa, D. Schmeidler, Updating ambiguous beliefs, Journal of Economic Theory 59 (1993) 33–49.
[27] H. Held, Quantile-filtered Bayesian learning for the correlation class, in: M. Zaffalon, G. de Cooman, J. Vejnarova (Eds.), ISIPTA, Action M Agency, 2007, pp. 223–232.
[28] H. Held, E. Kriegler, T. Augustin. Bayesian learning for a class of priors with prescribed marginals, vol. 488, 2006. Preprint server <http://www.stat.uni-muenchen.de/sfb386/>.
[29] H. Held, T. Schneider von Deimling, Transformation of possibility functions in a climate model of intermediate complexity, Advances in Soft Computing 6 (2006) 337–345.
[30] George J. Klir, Mark J. Wierman, Uncertainty-based information. Elements of generalized information theory, Physica (1999).
[31] E. Kriegler, H. Held, Utilizing random sets for the estimation of future climate change, International Journal of Approximate Reasoning 39 (2005) 185–209.
[32] M. Lavine, L. Wasserman, R.L. Wolpert, Bayesian inference with specified prior marginals, Journal of the American Statistical Association 86 (1991) 964–971.
[33] M. Martel-Escobar, F.J. Vázquez-Polo, A. Hernández-Bastida, Analysing the independence hypothesis in models for rare errors: an application to auditing, Applied Statistics 54 (4) (2005) 795–804.
[34] E. Moreno, Global bayesian robustness for some classes of prior distributions, in: D. Rios Insua, F. Ruggeri (Eds.), Robust Bayesian Analysis, Lecture Notes in Statistics, vol. 125, Springer, New York, 2000, pp. 45–70.
[35] R. Nelson, An introduction to copulas, Lecture Notes in Statistics, vol. 139, Springer, New York, 1998.
[36] R.C. Pacanowski, S.M. Griffies, MOM-3 manual, NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, USA, 1998.
[37] V. Petoukhov, A. Ganopolski, V. Brovkin, M. Claussen, A. Eliseev, C. Kubatzki, S. Rahmstorf, CLIMBER-2: a climate system model of intermediate complexity. Part i: model description and performance for present climate, Climate Dynamics 16 (2000) 1.
[38] D. Rios Insua, F. Ruggeri (Eds.), Robust Bayesian Analysis, Lecture Notes in Statistics, vol. 152, Springer, New York, 2000.
[39] H.V. Roberts, Probabilistic prediction, Journal of the American Statistical Association 60 (1965) 50–62.
[40] T. Seidenfeld, L. Wasserman, Dilation for convex sets of probabilities, Annals of Statistics 21 (1993) 1139–1154.
[41] S. Sivaganesan, A likelihood based robust Bayesian summary, Statistics and Probability Letters 43 (1999) 5–12.
[42] A.H. Tchen, Inequalities for distributions with given marginals, The Annals of Probability 8 (1980) 814–827.
[43] L. Tomassini, P. Reichert, R. Knutti, T.F. Stocker, M.E. Borsuk, Robust Bayesian uncertainty analysis of climate system properties using Markov chain monte carlo methods, Journal of Climate 20 (2007) 1239–1245.
[44] H.v. Storch, F.W. Zwiers, Statistical Analysis in Climate Research, Cambridge University Press, 1999.

[45] P. Walley, Statistical Reasoning with Imprecise Probabilities, Chapman and Hall, London, 1991.
[46] P. Walley, A bounded derivative model for prior ignorance about a real-valued parameter, Scandinavian Journal of Statistics 24 (1997) 463–483.
[47] P. Walley, Towards a unified theory of imprecise probability, International Journal of Approximate Reasoning 24 (2–3) (2000) 125–148.
[48] P. Walley, T.L. Fine, Towards a frequentist theory of upper and lower probability, The Annals of Statistics 10 (1982) 741–761.
[49] K. Weichselberger, The theory of interval-probability as a unifying concept for uncertainty, International Journal of Approximate Reasoning 24 (2000) 149–170.
[50] K. Weichselberger, Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als umfassendes Konzept, Physika, Heidelberg, 2001.
[51] K. Weichselberger, T. Augustin, On the competition and symbiosis of two concepts of conditional interval probability, in: Bernard et al. [7], pp. 608–629.
[52] C. Yu, F. Arasta, On conditional belief functions, International Journal of Approximate Reasoning 10 (1994) 155–172.